model, and show that it learns a feature set with quality comparable to a manually selected set for German. And for the second challenge we present results showing that it is possible to bridge the gap between a model trained on a predicted and another model trained on a projected morphologically enriched parallel corpus. Finally we exhibit final translation results showing promising improvement over the baseline phrase-based system.

Machine Translation-based Language Model Adaptation for Automatic Speech Recognition of Spoken Translations

Joris Pelemans, Tom Vanallemeersch, Kris Demuynck, Lyan Verwimp, Hugo Van Hamme, Patrick Wambacq

We present the results of our work on the SCATE - Smart Computer-Aided Translation Environment - project which addresses the integration of machine translation and automatic speech recognition for the recognition of spoken translations. We propose a technique that applies language model adaptation on the sentence level based on translation model probabilities of the source language text. We show that omitting language model renormalization after adaptation, as well as applying probability weights, drastically improves the efficiency compared to a similar technique described in the literature. Disk storage per sentence is reduced by ca. 5GB for a 3-gram language model and up to 15GB for a 5-gram language model. The time needed for adaptation takes ca. 0.2s per sentence which enables the integration of the technique into existing translation environments.

The effect on recognition accuracy is investigated for both word-based and phrased-based translation models and is combined with tailored models for named entities. The final model achieves a 25.3% relative error reduction compared to a 3-gram baseline without adaptation on a corpus of 167 English-to-Dutch spoken translations.

Mapping Leiden: Automatically extracting street names from digitized newspaper articles

Kim Groeneveld, Menno van Zaanen

The Dutch institute "Erfgoed Leiden en Omstreken" (ELO) has developed an interactive map of Leiden that allows people to search for monumental buildings and other interesting geographical entities or properties. We aim to extend this map with functionality that allows users to select a geographical area on the map and search for newspaper articles related to that area. This functionality requires an identification of street names in the collection of digitized newspaper articles. This task falls in the area of Geographical Entity Recognition (GER), which is the subfield of Named Entity Recognition (NER) that identifies geographical entities such as countries, cities, or street names.

In this research we evaluate two existing tools in the context of identifying street names: Memory Based Tagger (MBT) and the Stanford Named Entity Recognizer (Stanford NER). Based on an automatically created and manually checked Gold Standard (GS) dataset, both taggers are trained and tested using 10CV. Two experiments are performed. Firstly, learning curves (increasing the amount of training data) are created that show that with more training data, the tools are better at removing false positives (non-street names tagged as street names). Secondly, both systems are trained on data from which annotations of certain street names are removed. This allows us to investigate whether the systems are able to identify street names that are not annotated as